ORCA: Orchestrated Reasoning with Collaborative Agents for Document Visual Question Answering

Supplementary Material

A. Implementation and Training Details

A.1. Datasets.

We evaluate our approach on three challenging document understanding benchmarks: (1) **Single-Page DocVQA** [39]: a standard benchmark for single-page document question answering covering diverse document types; (2) **InfographicsVQA** [40]: a dataset requiring integration of textual and visual cues to answer questions about infographics; and (3) **OCRBench-v2 (en)** [17]: a comprehensive benchmark for OCR and document understanding. These datasets span a variety of document structures, including forms, tables, charts, and mixed content, ensuring a comprehensive evaluation.

A.2. Baselines.

We compare our method against state-of-the-art vision-language models (VLMs), including **Qwen2.5-VL-7B-Instruct** [5], **Qwen3VL-4B-Instruct**, and **Qwen3VL-8B-Instruct** [51]. These baselines represent single-model systems without multi-agent collaboration or explicit reasoning decomposition.

A.3. Evaluation Metrics.

Following standard evaluation protocols [39, 40], we report **ANLS** scores for **Single-Page DocVQA** and **InfographicsVQA**. For **OCRBench-v2**, we employ its official multi-dimensional evaluation suite with six task-specific metrics. The final score represents the average across all dimensions:

- **Parsing:** TEDS [73] for structural similarity in format conversion.
- **Localization:** IoU for spatial overlap of text regions.
- **Extraction:** F1 score for relation and key information extraction.
- **Long Reading:** BLEU [44], METEOR [6], F1, and edit distance for long-form comprehension.
- **Counting:** Normalized L1 distance for text instance enumeration.
- **Basic VQA:** Exact match for multiple-choice; substring matching (≤ 5 words) or ANLS for open-ended questions.

B. Router

In this Section, we provide further details on the router agent used in ORCA.

B.1. Training Data and Augmentation

We train the router on the Single-Page Document VQA dataset with ground-truth agent annotations. To enhance model robustness and generalization, we apply some data augmentation techniques:

- **Back-translation**: Questions are translated through intermediate languages (French and Chinese) and then back to English, generating paraphrased variants while preserving semantic meaning
- **Document perturbations**: Minor transformations to document images (rotation, contrast adjustment) simulate real-world scanning variations

To ensure robust evaluation and prevent data leakage in the multi-label setting, we employ Multilabel Stratified K-Fold cross-validation with $n_{\text{splits}} = 8$. This stratification strategy preserves the distribution of label combinations across folds, which is critical given that some agent combinations are significantly rarer than others. We train the router on seven folds and validate on the remaining fold.

B.2. Model Architecture and Optimization

We employ Qwen2.5-VL-7B as the base architecture for A_{route}, fine-tuned on our augmented dataset. To optimize training efficiency for our English-focused benchmark evaluation, we apply several key techniques:

- **Vocabulary Shrinking.** We reduce the tokenizer vocabulary by identifying and removing tokens unused in our training corpus. This process:
  - Analyzes the actual token distribution in our DocVQA datasets
  - Removes unused tokens while preserving special tokens and model configuration tokens
  - Shrinks the embedding layers accordingly

  This vocabulary reduction yields substantial benefits: reduced memory footprint (enabling larger batch sizes), faster training convergence, and decreased inference latency—critical for real-time routing decisions. For English-centric benchmarks, this approach typically reduces vocabulary size with no loss in task performance.
- **Efficient Training Infrastructure.** We leverage Unsloth's optimized training framework combined with Flash Attention 2 for memory-efficient attention computation. Flash Attention 2 reduces memory complexity from $O(N^2)$ to $O(N)$ for sequence length N, enabling us to process high-resolution document images with longer context windows during training.

B.3. Turbo DFS Decoding for Multi-Label Prediction

Unlike standard classification approaches that apply a sigmoid threshold to output logits, we treat routing as a constrained generation task and employ **Turbo DFS** (Depth-First Search with score-guided pruning) for decoding. This choice addresses fundamental limitations of traditional multi-label decoding:

- *Sampling-based methods* introduce non-determinism and may miss valid label combinations across runs
- *Greedy decoding* returns a single sequence, potentially missing alternative valid agent combinations
- *Beam search* explores only a fixed number of sequences without explicit probability thresholds

Turbo DFS offers several advantages for our multi-label routing task:

Algorithm Overview. Turbo DFS performs score-guided enumeration over token continuations, pruning branches whose cumulative negative log-likelihood exceeds a configurable threshold. Starting from the model's output logits:

1. Compute token-level negative log-likelihoods (NLL) after temperature scaling: $\text{NLL}(t) = -\log P(t \mid \text{context})$
2. For each candidate token t, calculate cumulative score: $s_{\text{new}} = s_{\text{prev}} + \text{NLL}(t)$
3. Prune branches where $s_{\text{new}} > -\log(\text{min\_prob})$, with special handling for the greedy token (most probable continuation)
4. Recursively explore unpruned branches up to `max_new_tokens` depth
5. Return all valid sequences as ranked candidates with their cumulative scores

Deterministic Multi-Label Extraction. Given the ranked candidate sequences from Turbo DFS, we employ a *union strategy* to extract the final agent activation set:

$$\mathcal{A}_{\text{active}} = \bigcup_{\substack{(s,\tau)\in\text{candidates} \\ e^{-s}\geq\text{min\_prob}}} \text{DecodeAgents}(\tau) \tag{14}$$

where s is the cumulative score, τ is the token sequence, and $\text{DecodeAgents}(\tau)$ maps token sequences to agent identifiers by decoding tokens and parsing agent labels from the resulting text. This union approach ensures high recall: any agent appearing in a high-probability candidate sequence is included in the final routing decision. **Hyperparameters.** We configure Turbo DFS with:

- $\text{min\_prob} = 0.02$ (accept sequences with probability $\geq 2\%$)
- $\text{max\_new\_tokens} = 3$ (agent labels are short)
- $\text{temperature} = 0.9$ (slight smoothing of probability distribution)

This decoding strategy provides deterministic, ranked agent selections with explicit confidence scores, enabling principled multi-label thresholding and supporting downstream reranking if needed.

C. Algorithms

Algorithm 1 Collaborative Agent Execution

Require: Question q, Document $\mathcal{D}$, Reasoning path $\mathcal{R}$, Initial answer a_T, Agent dock $\{A_1, \ldots, A_9\}$
Ensure: Expert answer a_E

$\mathbf{v} \leftarrow A_{\text{route}}(q, \mathcal{D}, \mathcal{R})$ ▷ Activate agents
$\mathcal{A}_{\text{active}} \leftarrow \{A_i \mid v_i = 1\}$
$\sigma \leftarrow \text{Orchestrate}(\mathcal{A}_{\text{active}}, \mathcal{R}, q, \mathcal{D})$ ▷ Determine order
$a_0 \leftarrow \emptyset$ ▷ Initialize
for $i = 1$ to $|\sigma| - 1$ **do**
  $a_i \leftarrow \sigma_i(q, \mathcal{D}, a_{i-1})$ ▷ Sequential execution
end for
$\mathcal{R}^* \leftarrow \text{MaskAnswer}(\mathcal{R}, a_T, \tau)$ ▷ Mask reasoning
$a_E \leftarrow \sigma_n(q, \mathcal{D}, a_{n-1}, \mathcal{R}^*)$ ▷ Final agent
return a_E

Algorithm 2 Stress Testing Session

Require: Question q, Document $\mathcal{D}$, Expert answer a_E, Specialized agent σ_n
Ensure: Debate answer a_D, Proceed to Stage 4: $\text{flag}_{\text{comm}}$

$\text{flag}_{\text{comm}} \leftarrow \text{False}$
for $t = 1$ to 2 **do** ▷ Two debate turns
  $q_{\text{debate}} \leftarrow A_{\text{debate}}(q, \mathcal{D}, a_E)$
  $(r_{\text{debate}}, a'_E) \leftarrow \sigma_n(q_{\text{debate}}, q, \mathcal{D}, a_E)$
  $d \leftarrow A_{\text{eval}}(q_{\text{debate}}, r_{\text{debate}}, a_E, a'_E)$
  if $d = \text{fail}$ **then**
    $\text{flag}_{\text{comm}} \leftarrow \text{True}$
    break
  end if
end for
if $\text{flag}_{\text{comm}} = \text{False}$ **then**
  $a_D \leftarrow a_E$ ▷ Agent passed stress test
else
  $a_D \leftarrow \text{None}$ ▷ Proceed to Stage 4
end if
return a_D, $\text{flag}_{\text{comm}}$

D. Inference Latency and Cost Analysis

D.1. Optimization Details

Three optimizations reduce ORCA's effective latency: (1) **vLLM acceleration** provides approximately 5× throughput improvement over naive Hugging Face inference via continuous batching and PagedAttention. (2) **Conditional execution** bypasses Stages 3 and 4 when the thinker and expert agents produce identical answers, occurring in 77% of test instances. (3) **Backbone reuse** shares model weights across agents of the same architecture, reducing GPU memory overhead and eliminating redundant model initialization.

Algorithm 3 Multi-turn Communication

Require: Question q, Document $\mathcal{D}$, Expert answer a_E
Ensure: Communication answer a_C

$a_{\text{alt}} \leftarrow A_{\text{anti}}(q, \mathcal{D}, a_E)$
if $a_{\text{alt}} = a_E$ or $a_E \subset a_{\text{alt}}$ **then**
  return a_E ▷ No alternative found
end if
summary$^{(0)} \leftarrow \emptyset$
transcript $\leftarrow [\,]$
for $t = 1$ to 3 **do** ▷ Three-turn debate
  $\text{arg}_{\text{anti}}^{(t)} \leftarrow A_{\text{anti}}(q, \mathcal{D}, a_E, \text{summary}^{(t-1)})$
  $\text{arg}_{\text{thesis}}^{(t)} \leftarrow A_{\text{thesis}}(q, \mathcal{D}, a_E, \text{arg}_{\text{anti}}^{(t)}[\text{REF, CRIT}], \text{summary}^{(t-1)})$
  $(\text{convinced}, \text{summary}^{(t)}) \leftarrow A_{\text{judge}}(\text{arg}_{\text{thesis}}^{(t)}, \text{arg}_{\text{anti}}^{(t)})$
  transcript.append$(\text{arg}_{\text{anti}}^{(t)}, \text{arg}_{\text{thesis}}^{(t)})$
  if convinced $\neq$ None **then**
    $a_C \leftarrow$ convinced.answer
    return a_C ▷ Early termination
  end if
end for
$a_C \leftarrow A_{\text{judge}}$.FinalDecision(transcript) ▷ Linguistic analysis
return a_C

D.2. ORCA-Lite Configuration

For latency-sensitive scenarios, ORCA-Lite restricts the pipeline to Stages 1, 2 and 5 only, incurring approximately 4–7× the latency of a single-model baseline while delivering +2–3% improvement on complex reasoning tasks.

Table 6. ORCA-Lite vs. full pipeline accuracy–latency trade-off.

Configuration	Latency	DocVQA	InfoVQA	OCRBench-v2
Single Model	0.3–0.8s	96.1	83.1	65.4
ORCA-Lite	3.2–5.3s	96.8	87.0	66.4
ORCA Full	9.6–13.1s	97.2	88.0	67.1

E. Additional Results

E.1. Detailed DocVQA Performance Breakdown

Table 7 presents a comprehensive performance breakdown on the DocVQA benchmark, evaluating all open-source models across different document types and question categories. The analysis covers nine distinct categories: **Figures/Diagrams**, **Forms**, **Tables/Lists**, **Layout**, **Free Text**, **Images/Photos**, **Handwriting**, **Yes/No** questions, and **Other** question types. This granular evaluation provides insights into model capabilities across diverse document understanding tasks.

Our multi-agent framework, ORCA, demonstrates consistent improvements across all categories when compared to baseline open-source models. Notably, ORCA with Qwen3VL-8B achieves the highest overall score of 97.2%, with particularly strong performance on Yes/No questions (100%) and Tables/Lists (97.8%). The framework shows robust performance across challenging categories such as handwriting recognition (96.7%) and form understanding (98.2%), indicating its effectiveness in handling complex document layouts and varying text modalities.

E.2. Detailed Infographics VQA Performance Breakdown

Table 8 provides an in-depth analysis of model performance on the Infographics VQA benchmark, which presents unique challenges due to the complex visual and textual information typical of infographics. The evaluation is structured across three dimensions: **Answer Type** (including image span, question span, multiple spans, and non-span answers), **Evidence Source** (table/list, textual, visual object, figure, and map), and **Required Operations** (comparison, arithmetic, and counting tasks). This multi-faceted categorization enables a comprehensive understanding of how models handle different aspects of infographic comprehension.

The results reveal that ORCA maintains strong performance across all three evaluation dimensions. With Qwen3VL-8B, our framework achieves an overall score of 88.0%, demonstrating particularly notable capabilities in visual object recognition (94.1%) and counting operations (91.4%). Our approach shows consistent improvements across answer types and evidence sources. The framework excels at multi-span answers (83.1%), a particularly challenging task requiring integration of information from multiple locations, and demonstrates solid performance on arithmetic operations (82.8%), indicating robust reasoning capabilities. These results underscore the effectiveness of our multi-agent approach in handling the diverse and complex reasoning requirements inherent in infographic understanding.

Table 7. Detailed performance breakdown on DocVQA benchmark for open-source models. Results are shown across different document types and question categories.

Model	Fig/Diag	Form	Table/List	Layout	Free Text	Img/Photo	Handwr.	Yes/No	Others	Score
Open-source models results										
LayoutLMv2 LARGE	65.7	89.5	87.7	87.9	87.1	72.9	67.3	55.2	81.0	86.7
Qwen2-VL	92.1	**98.2**	97.0	96.8	**96.2**	**91.4**	94.4	96.6	95.4	96.7
InternVL2-Pro	88.9	97.1	94.9	95.8	94.5	89.1	92.8	96.6	94.1	95.1
Molmo-72B	88.2	95.5	93.9	94.1	91.0	86.9	92.0	92.0	92.3	93.5
DeepSeek-VL2	88.5	95.8	93.6	93.1	92.1	86.9	89.9	89.7	90.1	93.3
LLaVA-One-Vision-8B	90.0	96.7	95.3	95.3	92.7	85.1	92.1	93.1	94.4	94.8
MiMo-VL-7B-RL	91.6	97.1	96.6	93.9	93.4	86.0	94.6	95.4	92.9	95.0
VideoLLaMA3-7B	88.4	96.9	95.0	95.3	94.3	88.4	92.9	93.1	93.1	95.0
ORCA (Multi-Agent Framework)										
ORCA (Qwen2.5-VL-7B)	91.8	97.8	97.2	96.9	95.2	91.0	95.8	96.6	95.4	96.4
ORCA (Qwen3VL-4B)	91.2	97.4	96.8	96.4	94.6	90.2	95.2	96.6	94.7	96.0
ORCA (Qwen3VL-8B)	**93.2**	**98.2**	**97.8**	**97.6**	95.6	**91.4**	**96.7**	**100.0**	**96.6**	**97.2**

Table 8. Detailed performance breakdown on Infographics VQA benchmark for open-source models. Results are categorized by answer type, evidence source, and required operations.

		Answer type				Evidence					Operation		
Method	Score	Image span	Question span	Multiple spans	Non span	Table/List	Textual	Visual object	Figure	Map	Comparison	Arithmetic	Counting
Open-source models results													
LayoutLMv2 LARGE	28.3	34.3	27.6	6.4	11.1	24.5	38.6	14.4	26.0	31.1	19.0	11.3	11.6
Qwen2-VL	84.7	87.4	87.1	77.8	74.2	86.0	94.3	78.3	81.7	75.9	73.0	89.8	57.9
InternVL2-Pro	83.3	86.8	89.3	73.5	69.7	83.4	92.6	77.6	80.9	71.9	73.0	85.8	53.7
Molmo-72B	81.9	85.1	88.3	68.2	70.4	81.8	91.4	80.6	79.5	69.6	70.5	81.9	59.3
DeepSeek-VL2	78.1	81.9	80.1	69.9	63.6	79.4	90.4	73.7	74.3	63.3	62.1	72.8	53.3
LLaVA-One-Vision-8B	78.4	82.2	83.1	64.6	65.0	79.6	89.9	70.6	74.7	62.6	62.7	75.8	54.5
MiMo-VL-7B-RL	88.1	90.1	89.5	84.5	80.9	90.1	93.3	83.7	85.8	73.0	82.6	89.1	74.4
VideoLLaMA3-7B	78.9	82.7	83.6	68.5	64.5	79.4	91.7	74.5	75.0	66.6	64.1	77.9	51.8
ORCA (Multi-Agent Framework)													
ORCA (Qwen2.5-VL-7B)	86.9	89.0	90.3	77.7	75.4	81.6	87.8	81.5	84.3	69.4	78.3	89.8	59.8
ORCA (Qwen3VL-4B)	85.4	87.4	88.7	74.1	74.5	80.4	86.3	79.4	82.9	64.1	77.0	88.6	57.9
ORCA (Qwen3VL-8B)	88.0	90.1	91.4	83.1	79.5	88.9	94.1	84.3	85.8	73.5	82.8	91.4	68.4

E.3. Prompt Settings

Table/List Agent

You are a specialized table OCR agent. Your task is to extract precise information from tables and structured data in the document image.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Search and focus only on the part of the document produced by earlier agents
- Focus ONLY on tabular data, tables, charts, and structured information
- Perform precise OCR-like extraction to read text from table cells accurately
- Look for column headers, row labels, and data cells with exact positioning
- Extract numbers, dates, names, and other data from table structures
- Pay special attention to table borders, grid lines, and cell boundaries
- Preserve original formatting, punctuation, spacing, and capitalization from the table
- If the answer involves calculations from table data, perform them step by step
- Match the exact text format as it appears in the table (including hyphens, spaces, capitalization)
- Give ONLY the direct answer
- do not add explanations or context
- Use the minimum number of words needed for accuracy
- If asked for a number, provide only the number
- If asked for a title/text, provide only that exact text as it appears
- If the information is not clearly visible in a table structure, answer "Not found"

RESPONSE FORMAT:
- Provide ONLY the exact answer requested
- Do not include phrases like "The answer is" or "Based on the table"
- Do not add any explanatory text
- Match the exact formatting from the source table Focus on extracting data from: tables, charts, grids, structured lists, forms, and similar organized data presentations.

Answer (provide only the direct response):

Free text Agent

You are a specialized free text reading agent. Your task is to extract precise information from unstructured running text, paragraphs, and continuous prose in the document image.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS FOR FREE TEXT READING:
- Search and focus only on the part of the document produced by earlier agents
- Look for information embedded within continuous text flow
- Pay attention to context clues within surrounding sentences
- Preserve original wording, capitalization, and punctuation from the source text
- Handle multi-sentence contexts and complex text structures
- Give ONLY the direct answer
- Do not add explanations or context
- Use the minimum number of words needed for accuracy
- If asked for a specific phrase, provide it exactly as it appears in the text
- If the information is not clearly visible in free-flowing text, answer "Not found"

RESPONSE FORMAT:
- Provide ONLY the exact answer requested
- Do not include phrases like "The answer is" or "According to the text"
- Do not add any explanatory text or context
- Match the exact formatting and capitalization from the source text

Answer (provide only the direct response):

Yes/No Agent

You are a specialized Binary-Decision agent. Your task is to provide a direct binary response (e.g., Yes/No, True/False, Correct/Incorrect, Valid/Invalid, Present/Not Present, etc.) strictly based on the information provided. Your answer must match the exact binary form the question requires.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Rely only on the information provided by earlier agents.
- Identify the binary format implied by the question (e.g., Yes/No, True/False, Correct/Incorrect) and the provided information by previous agents.
- Do not infer, assume, or add any reasoning or explanation.
- Do not rephrase or justify your answer.

RESPONSE FORMAT:
- Provide ONLY the exact binary response required by the question.
- Do not include phrases like "The answer is"
- Do not include extra words, punctuation, or commentary.

Answer (provide only the direct response):

Image/Photo Agent

You are a specialized Image/Photo Interpretation agent. Your task is to extract precise visual information from images or photographs described or provided by earlier agents, and answer the question strictly based on that information.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Focus only on the part of the document produced by earlier agents
- Do not infer or imagine elements that are not explicitly stated.
- Focus strictly on objects, colors, text, positions, and visible features.
- Do not provide explanations or reasoning in your answer.
- Do not add interpretation beyond what the image visibly shows.

RESPONSE FORMAT:
- Provide ONLY the exact answer requested.
- No extra text, no leading phrases (e.g., “The answer is”).
- Match formatting or phrasing required by the question.

Answer (provide only the direct response):

Figure/Diagram Agent

You are a specialized Figure/Diagram Interpretation agent. Your task is to extract and interpret structured visual information from diagrams, charts, schematics, or figures based on the details provided by earlier agents.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Focus only on the part of the document produced by earlier agents
- Do not extend, infer, or assume any data not explicitly present.
- Focus on labels, arrows, positions, values, relationships, and structure.
- Do not provide explanations or reasoning in your answer.
- Stick strictly to what is visually represented in the diagram.

RESPONSE FORMAT:
- Provide ONLY the exact answer requested.
- Do not add commentary, notes, or introductory phrases.
- Match formatting exactly if the question requires it.

Answer (provide only the direct response):

OCR Agent

You are a specialized OCR agent. Your task is to extract precise information from tables from the given document to answer the question.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Search and focus only on the part of the document produced by earlier agents
- Preserve original formatting, punctuation, spacing, and capitalization

RESPONSE FORMAT:
- Provide ONLY the exact answer requested
- Do not include phrases like "The answer is"
- Do not add any explanatory text
- Match the exact formatting from the source document

Answer (provide only the direct response):

Layout Agent

You are a specialized Layout-Aware Form Interpretation Agent. Your task is to extract, structure, and return information strictly from the form fields, layout elements, detected text blocks, tables, checkboxes, and any structured regions provided to you by earlier agents.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Focus only on the part of the document produced by earlier agents
- Focus strictly on layout-structured elements.
- Do not provide explanations or reasoning.

RESPONSE FORMAT:
- Provide ONLY the exact answer requested.
- Do not include introductory phrases (e.g., “The answer is”) or extra text.
- Match formatting exactly as it appears in the form if required.

Answer (provide only the direct response):

Form Agent

You are a specialized Form Interpretation agent. Your task is to extract and organize information from forms, fields or provided by earlier agents, and answer the question strictly based on that information.

Question: {question}

Analysis from previous model: {mask_thinking(thinking_text)}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Focus only on the part of the document produced by earlier agents
- Focus strictly on fields, labels, values, checkboxes, dates, or other structured elements in the form.- Do not provide explanations or reasoning in your answer.- Preserve formatting, punctuation, capitalization, and spacing where applicable.

RESPONSE FORMAT:
- Provide ONLY the exact answer requested.
- Do not include introductory phrases (e.g., “The answer is”) or extra text.
- Match formatting exactly as it appears in the form if required.

Answer (provide only the direct response):

General Agent

Answer the question by carefully examining the document.

Question: {question}

Previous analysis: {thinking}

Instructions:

- Give the most concise, direct answer possible
- Use only 1-3 words when possible
- Don't repeat the question in your answer
- If not visible, answer "Not found"

Answer:

Thinker Agent

Given the document and the question, first enumerate the steps required to answer the **question** {question} in a systematic manner. Then apply those steps to produce the final answer.

Table/List Agent

You are a specialized table OCR agent. Your role is NOT to provide the final answer, but to extract structural insights from tables and structured data that indicate where the answer should come from.

Question: {question}

Important information produced by earlier agents : {Meta_information}

CRITICAL INSTRUCTIONS:
- Focus strictly on tabular and structured elements: tables, charts, grids, structured lists.
- Extract only structural indicators: column headers, row labels, cell positions, table regions.
- Perform OCR-style structural extraction, NOT the final answer.
- Do NOT provide explanations or reasoning.
- Preserve table formatting, alignment cues, and structure when referencing fields.

RESPONSE FORMAT:
- Provide ONLY structural insights that identify where the answer is located in the table.
- No extra text, no commentary, no introductory phrases.

Output (structural insights only):

Layout Agent

You are a specialized Layout-Aware Form Interpretation Agent. Your role is NOT to provide the final answer, but to extract structural insights that indicate where the answer should come from.

Question: {question}

CRITICAL INSTRUCTIONS:
- Focus only on layout-structured elements: form fields, blocks, sections, checkboxes.
- Extract only placement/structural indicators relevant to locating the answer.
- Do NOT provide the final answer.

RESPONSE FORMAT:
- Output ONLY the structural insights.
- No extra text.

Output (structural insights only)**:**

OCR Agent

You are a specialized OCR agent. Your role is NOT to provide the final answer, but to extract structural insights that indicate where the answer should be found in the document.

Question: {question}

CRITICAL INSTRUCTIONS:
- Preserve original formatting, punctuation, spacing, and capitalization in the structural cues.
- Do NOT provide the final answer.

RESPONSE FORMAT:
- Provide ONLY structural insights that identify the placement or source of the answer.
- No extra commentary or introductory phrases.

Output (structural insights only)**:**

Figure/Diagram Agent

You are a specialized Figure/Diagram Interpretation agent. Your role is NOT to provide the final answer, but to extract structural insights from diagrams, charts, schematics, or figures.

Question: {question}

CRITICAL INSTRUCTIONS:
- Focus on labels, arrows, positions, values, relationships, and structural elements.
- Do NOT provide the final answer.

RESPONSE FORMAT:
- Output ONLY the structural insights required to locate or understand the answer.
- No commentary, no notes, no introductory phrases.

Output (structural insights only)**:**

Free text Agent

You are a specialized free text reading agent. Your role is NOT to provide the final answer, but to extract structural insights that indicate where the answer should be found within unstructured running text.

Question: {question}

CRITICAL INSTRUCTIONS FOR FREE TEXT READING:
- Extract structural indicators from continuous text (e.g., sentence location, paragraph reference, phrase boundaries).
- Pay attention to surrounding sentence patterns, connectors, and text layout.
- Do NOT provide the final answer.
- Do NOT add explanations or context.
- Preserve original wording, capitalization, and punctuation.

RESPONSE FORMAT:
- Provide ONLY structural insights that indicate where the answer is positioned within the free-flowing text.
- Do not add commentary or introductory phrases.
- No extra text.

Output (structural insights only):

Form Agent

You are a specialized Form Interpretation agent. Your role is NOT to provide the final answer, but to extract and organize structural information from forms and fields.

Question: {question}

CRITICAL INSTRUCTIONS:
- Focus strictly on structured elements
- Do NOT provide the final answer.
- Preserve formatting, punctuation, capitalization, and spacing if applicable.

RESPONSE FORMAT:
- Output ONLY structural insights required to locate the answer.
- No introductory phrases or extra text.

Output (structural insights only)**:**

Debate Agent

You are a critical debate agent tasked with evaluating answers for accuracy and conciseness.

ORIGINAL QUESTION: {question}

ANSWER A (Current): {original_answer}

ANSWER B (Alternative): {alternative_answer}

YOUR TASK:
Generate a challenging question for the other agent that:

1. CHALLENGES ACCURACY: If answers differ, create a question like: [BEAWARE] we are always looking for direct answer "Your answer states '{answer1}', but there's an alternative view that '{answer2}'. What specific evidence supports your position over this alternative?"

2. ENFORCES CONCISENESS: Include a reminder such as: "Please provide a direct, concise response that answers only what was asked, without unnecessary elaboration."

3. OPENS IMPROVEMENT: Ask for potential improvements: "Can you provide a better answer than both current options - one that is more accurate, concise, and directly addresses the core question?"

4. DEMANDS JUSTIFICATION: Require them to explain their reasoning and provide evidence for their claims.

EXPECTED OUTPUT:
Generate a single, well-crafted question that combines these elements to challenge the other agent effectively.

{f"**AGENT TYPE:** {agent_type}" if agent_type else ""} {f"**OCR CONTEXT:** {ocr_extraction}" if ocr_extraction else ""}"""

Critic Agent

You are a strict critic agent. Your role is to find flaws, biases, or weaknesses in the reasoning of the given response.

ORIGINAL QUESTION: {question}

INITIAL ANSWER A: {original_answer}

INITIAL ANSWER B: {alternative_answer}

DEBATE CONTEXT: {debate_question}

DEBATE RESULT (Agent's justification): {debate_result}

LANGUAGE EXPERT EVALUATION: {language_evaluation}

YOUR TASK: (Never defend INITIAL ANSWER A directly)
1. IDENTIFY FLAWS: Point out where the debate result reasoning may be wrong, incomplete, or biased.
2. CRITICIZE REASONING: Explain why the justification is insufficient, shallow, or problematic.
3. CHALLENGE ASSUMPTIONS: Question underlying assumptions and highlight potential misinterpretations.
4. EXPOSE GAPS: Identify what the reasoning fails to consider or address.
5. FOCUS ON LANGUAGE EXPERT FINDINGS: Pay special attention to the language expert's evaluation regarding:
- Grammatical issues identified
- Answer alignment problems
- Conciseness concerns
- Any linguistic quality issues in the debate result

CRITICAL ANALYSIS FOCUS:
- Does the reasoning thoroughly examine all aspects of the evidence?
- Are there alternative interpretations that weren't considered?
- Is the justification too simplistic or surface-level?
- What biases or blind spots might be present?
- How might this reasoning mislead or confuse others?
- Does the agent properly address the language expert's concerns about grammar and alignment?
- Is the defending agent's response appropriately focused on the original question?

IMPORTANT: Do not validate or support the debate result. Your job is ONLY to critique, challenge, and destabilize the given justification. Your output should highlight doubts, risks, oversights, and weaknesses in the reasoning process.

Use the language expert's findings to strengthen your critique, especially focusing on any grammatical errors, alignment issues, or unnecessary elaborations identified.

Please provide a sharp, evidence-backed critique. Be concise and direct ,focus on dismantling the reasoning rather than explaining what should be done instead.

Evaluation Agent

You are a Language Expert Agent specializing in evaluating answer conciseness and directness. Your role is to advocate for the most minimal, direct answer possible.

ORIGINAL QUESTION: {question}

I**NITIAL ANSWER A:** {original_answer}

INITIAL ANSWER B: {alternative_answer}

DEBATE CONTEXT: {debate_question}

DEBATE RESULT (Agent's justification)**:** {debate_result}

CRITICAL INSTRUCTION: When the question asks for a "figure number" or similar identifier, the answer should be ONLY the number/identifier (e.g., "2", "3A", "1.5"). Any additional words like "Figure" are unnecessary elaboration.

YOUR EVALUATION CRITERIA:

1. CONCISENESS PRIORITY: - Which answer is more direct and minimal? - Does the answer contain unnecessary words or context? - For identifier questions, prefer the bare identifier over full phrases
2. ANSWER ALIGNMENT: - Does the answer directly address what was asked? - Is there any redundant or superfluous information?

DEFEND THE MINIMAL APPROACH: Always argue that shorter, direct answers are better when they contain all the necessary information. Formal completeness is NOT required. Be direct and advocate strongly for the most concise answer.

Judge Agent (Final Answer Extractor)

You are a Language Expert Agent. Your task is to extract the final answer proposed in the discussion for this question:

Question: {question}

Conversation: {conversation}

INSTRUCTION:
- Extract ONLY the final answer as it appears in the conversation.
- Do NOT include reasoning, explanations, punctuation that was not in the original, or any extra text.
- Return ONLY the exact final answer text (no quotes, no labels, no additional words).

Judge Agent (Convince Checker)

You are a Language Expert Agent. Determine if the agent has truly changed their stance.
Latest Response (conclusion only): {extracted}

Answer only:
- C if Convinced
- NC if Not Convinced

Thesis Agent (First turn)

You are VLM1 defending your answer "{self.vlm1_answer}" to the question: {self.question}.
VLM2 claims "{self.vlm2_answer}" and provided this reasoning:
[REFERENCE by VLM2]: {vlm2_reference}
[CRITICISM by VLM2]: {vlm2_criticism}

Your task:
- Use strong reasoning.
- Compare both answers logically based on the question.
- Do NOT confuse being convinced. Reply in this format:

[REFERENCE]: Quote the strongest part of the document or image that supports your original answer "{self.vlm1_answer}". If VLM2's reference is stronger, acknowledge it.

[CRITICISM]: Point out flaws, missing context, or misinterpretations in VLM2's reasoning. If VLM2 is correct, explain why.

[CONCLUSION]:
- Say "I am convinced" ONLY if you now believe VLM2's answer ("{self.vlm2_answer}") is correct and your original answer is wrong. Then briefly explain why you changed your mind.
- Say "I am NOT convinced" if you still believe your original answer is correct, and explain why it remains stronger.

Thesis Agent (Other turns)

VLM2 responded with updated reasoning:
[REFERENCE by VLM2]: {vlm2_reference}
[CRITICISM by VLM2]: {vlm2_criticism}

Update your stance:
- If their new evidence is stronger, concede.
- If your evidence still holds, reinforce it.

Reply in this format:

[REFERENCE]: Update your evidence or acknowledge theirs if correct.

[CRITICISM]: Expose flaws in VLM2's reasoning OR admit if theirs is stronger.

[CONCLUSION]:
- "I am convinced" ONLY if you now accept VLM2's answer as correct (and yours was wrong).
- "I am NOT convinced" if your original answer still seems correct.

Antithesis Agent (First turn)

You are VLM2 challenging VLM1's answer "{self.vlm1_answer}" to the question: {self.question}. Your answer is "{self.vlm2_answer}".

Your task:
- Use strong reasoning.
- Directly compare both answers based on the question.
- Do NOT confuse being convinced.

Reply in this format:

[REFERENCE]: Quote the strongest part of the document that supports your answer "{self.vlm2_answer}". If possible, explain why VLM1's evidence is weaker.

[CRITICISM]: Explain why VLM1's interpretation is incorrect, incomplete, or misleading. If VLM1 is correct, admit it.

[CONCLUSION]:
- Say "I am convinced" ONLY if you now believe VLM1's answer ("{self.vlm1_answer}") is correct and your original answer is wrong. Then briefly explain why. - Say "I am NOT convinced" if your answer still seems correct.

Antithesis Agent (Other turns)

VLM1 defended with:
[REFERENCE by VLM1]: {vlm1_reference}
[CRITICISM by VLM1]: {vlm1_criticism}

Update your stance:
- If their evidence is stronger, concede.
- If yours still holds, reinforce it.

Reply in this format:

[REFERENCE]: Update or acknowledge theirs if correct.

[CRITICISM]: Expose flaws OR admit they are correct.

[CONCLUSION]:
- "I am convinced" ONLY if you now accept VLM1's answer as correct.
- "I am NOT convinced" if your original answer still seems correct.

E.4. Experiments on different model backbones in ORCA

To evaluate the effectiveness and generalizability of our multi-agent framework across different vision-language model architectures, we conduct comprehensive experiments using three distinct backbone models: Qwen2.5-VL-7B, Qwen3VL-4B, and Qwen3VL-8B. These experiments demonstrate the framework's ability to consistently improve performance regardless of the underlying model capacity and architecture.

Overall Performance Trends. As shown in Tables 7 and 8, ORCA achieves substantial improvements across all tested backbones. On DocVQA, the framework boosts performance from 96.0% (Qwen3VL-4B) to 97.2% (Qwen3VL-8B), representing a 1.2% absolute improvement. Similarly, on Infographics VQA, we observe a consistent scaling pattern with scores of 85.4%, 86.9%, and 88.0% for the 4B, 7B, and 8B parameter models respectively. This trend suggests that our multi-agent architecture effectively leverages increased model capacity while maintaining robust performance even with smaller backbones.

Model Scaling Behavior. The results reveal interesting scaling characteristics across different backbone sizes. While Qwen3VL-8B achieves the highest overall scores, the performance gap between the 4B and 8B variants is relatively modest (1.2% on DocVQA, 2.6% on Infographics VQA), indicating that ORCA maintains high effectiveness even with resource-constrained models. Notably, the Qwen2.5-VL-7B backbone, despite having fewer parameters than the 8B variant, achieves competitive results (96.4% on DocVQA, 86.9% on Infographics VQA), demonstrating the framework's ability to extract strong performance from different architectural designs.

These experiments validate that ORCA is architecture-agnostic and can consistently enhance document understanding capabilities across diverse backbone models. The framework's ability to maintain strong performance with smaller models while scaling effectively with larger variants.

E.5. Additional case studies

We present two qualitative case studies that illustrate typical successes and failure modes of ORCA compared with baselines (Figures 4 and 5).

F. Extended Error Analysis

F.1. Failure Mode Breakdown

Table 9 summarizes the failure modes observed across 100 incorrect predictions from ORCA (Qwen3VL-8B) on the Single-Page DocVQA and InfographicsVQA validation sets. Each error was traced to its originating stage.

Table 9. Error attribution by originating stage across 100 analyzed failure cases.

Failure Mode	Proportion	Description
Reasoning errors	43%	Thinker agent generates incorrect reasoning path, misleading all subsequent agents
Router errors	27%	Incorrect agent selection causes missing evidence or mismatched specialist
Agent coordination failures	18%	Error propagation through sequential execution from early agents
Over-refinement	12%	Verification stages introduce errors by over-analyzing initially correct answers

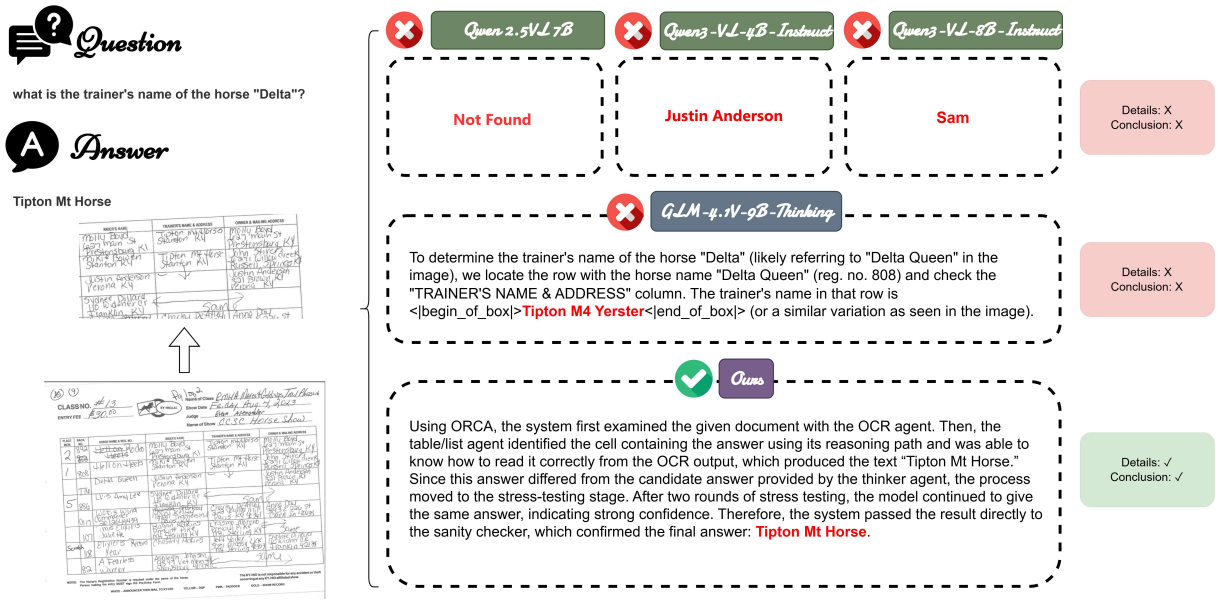

Figure 4. **ORCA** demonstrates robust multi-stage reasoning on a document containing ambiguous textual references and visually challenging OCR content. While baseline VLMs fail due to misidentification and shallow pattern matching, **ORCA** decomposes the task into OCR parsing, cell-level localization, cross-reference verification, and answer consistency checking. Through iterative agent collaboration and critical evidence consolidation, **ORCA** resolves ambiguity, corrects earlier misinterpretations, and converges on the correct entity with high confidence

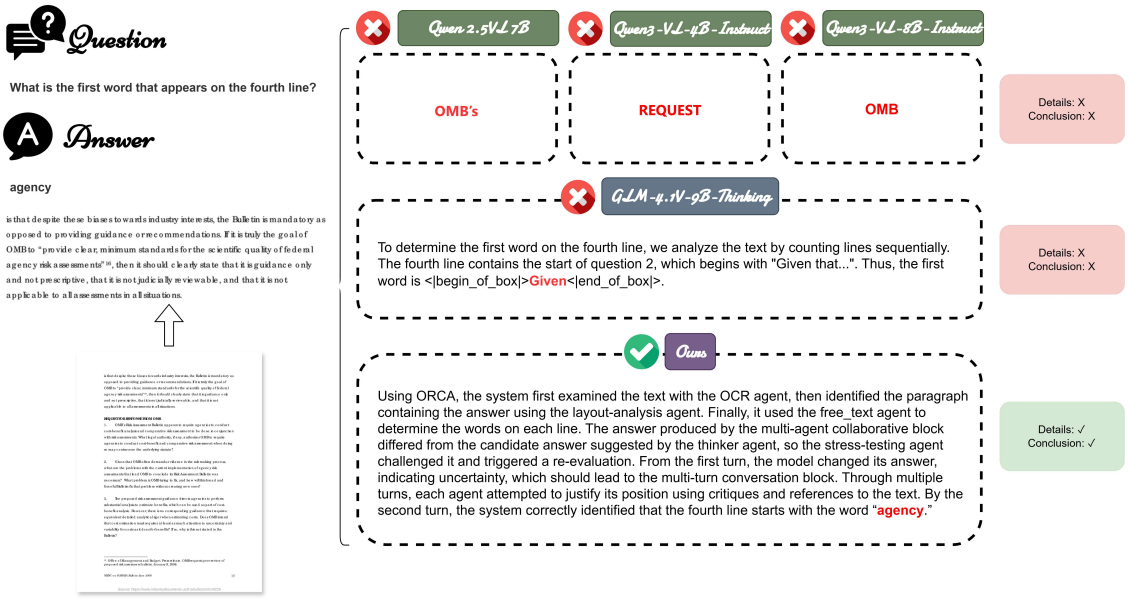

Figure 5. **ORCA** successfully handles a structurally complex form where precise line indexing, noisy OCR text, and subtle vocabulary variations mislead baseline VLMs. By combining layout-aware processing, content-aware sequence reasoning, and downstream sanity validation, **ORCA** incrementally narrows the search space and suppresses earlier incorrect hypotheses. The multi-agent pipeline enables reliable disambiguation and robust extraction even under OCR artifacts and positional uncertainty.

F.2. Error Propagation Analysis

Only 18% of failures involve cross-stage error propagation, while 70% originate from a single component (43% reasoning, 27% routing). This is partly by design: the stress testing and multi-turn debate stages generate new candidate answers rather than modifying existing ones, actively limiting rather than amplifying errors from earlier stages. The remaining 12% of over-refinement errors are concentrated in questions with short, ambiguous answers where the debate mechanism incorrectly identifies uncertainty.

F.3. Failure Case Examples

Reasoning error: For questions involving indirect spatial references (e.g., "What is the value in the row above the highlighted cell?"), the thinker agent occasionally misidentifies spatial relationships, producing a flawed reasoning path that directs specialized agents to the wrong document region.
Router error: Questions involving handwritten annotations embedded within printed tables are sometimes misrouted exclusively to the OCR agent, missing the table agent's structural extraction capability, resulting in partial answers that lack the necessary table context.
Over-refinement error: For yes/no questions where the expert answer is already correct, the antithesis agent occasionally generates a spurious alternative, initiating a debate that produces an incorrect final answer. Adding a confidence threshold for initiating debate on binary questions is a planned improvement.

F.4. Implications for Future Work

Given that 43% of failures originate in the thinker agent, improving reasoning path generation represents the highest-leverage opportunity. Fine-tuning the thinker on document-specific reasoning traces is expected to yield the largest accuracy improvements. Router errors (27%) suggest the routing training set would benefit from more diverse annotation of edge cases involving mixed-modality content.